An Imaging System Correlating Lip Shapes with Tongue Contact Patterns for Speech Pathology Research

D. J. Lee [a], Daniel Bates [a], Christopher Dromey [b], Xiaoqian Xu [a], and Sameer Antani[c]
[a] *Dept. of Electrical and Comp. Eng., Brigham Young University, Provo, Utah*
[b] *Dept. of Audiology and Speech-Language Pathology, Brigham Young University*
[c] *National Library of Medicine, Bethesda, Maryland*
djlee@ee.byu.edu, dmb677@yahoo.com, cdd32@email.byu.edu, xiaoqian@et.byu.edu, and antani@lhc.nlm.nih.gov

Abstract

In this research, an imaging system was built to work with a newly developed electronic device to help people produce sounds correctly. The system consists of two parts, the internal tongue contact pattern data collection and the external lip shape information analysis. The tongue position information was gathered using the palatometer, an innovative tongue contact pattern-tracking device invented by Dr. Samuel Fletcher. The lip shape information was collected by processing images taken from people articulating different sounds. We developed an efficient color image segmentation technique to extract lip contour points and form a closed curve for shape analysis. The geometry invariant turn function vs. normalized length was then calculated for the lip shape for each sound and compared against the turn function of the lips in a resting position to quantify their variations from this reference. Both internal (vocal tract) and external (visible lip shape) information was collected for each of the speech sounds. The lip shape information extracted from the images was then correlated with tongue position information. The test results showed that this imaging system can be used to quantify the lip shape information and its relations with the tongue position and is a potentially useful tool for speech pathology research.

1. Introduction

Many speech readers are able to interpret speech by observing the lip movements and facial expressions of the speaker. This demonstrates that there is information conveyed visually during the process of speech allowing recognition of what is being said. Even those with normal hearing are able to better understand what is being said if they are able to see the face of the person speaking, especially under noisy conditions [1]. In fact, while listening, so much information is taken visually from speech that the mismatch between auditory and visual signals can lead to perceptual illusions. In the McGurk effect [2], listeners' perception of speech sounds is different when they see a video image of the speaker whose facial movements and speech are incongruous. Typically, when listeners hear *baba* but see *gaga*, they perceive *dada*. Listeners are usually surprised to perceive *baba* when they look away from the video screen. It is therefore clear that the visual signals play a significant role in everyday auditory-verbal communication.

These observations have motivated some to intertwine lip-reading with audio signals in computer speech recognition systems. Petajan used the current technology of dynamic time-warping with visual features derived from mouth opening and demonstrated that an audio-visual system was more effective in recognizing speech than either speech or vision alone [3].

Bregler and Konig have also shown through their experiments that the incorporation of additional visual information can significantly reduce recognition error [4]. They created scenarios that degraded the quality of the speech signal in such a way that state of the art speech recognition systems were incapable of good recognition performance. They simulated car noise and cross talk and added it to the clean speech signal at different ratios.

Chen and Rao reviewed recent research that examines audio-visual integration in multimodal communication [5]. They argued that speech communication in its nature is conveyed by more than the just auditory channel. A good example of this is the sounds /p/ and /k/ which can easily be distinguished by the visual cue of the open or closed mouth. They proposed that analogous to the phoneme in the acoustic domain is 'viseme' in the visual domain. The viseme is the smallest visibly distinguishable unit of speech. For correct articulation, the shape the mouth makes during the production of the viseme is important. In this research we used /s/, /sh/ and /m/ visemes along with /a/, /e/, /i/, /o/, and a reference image and correlated them with their tongue contact patterns recorded during sound production. The correlation between the two enhances our understanding of multi-articulator coordination in speech production.

The novelties of this work are the use of a modified shape representation method to quantify the shape difference from a reference shape and to use this quantified difference to determine the relationship between vocal tract movements and externally visible lip shape during sound production. Shapes were described more accurately in this way than merely using width and height of a few control points as has been done in the past. This research was not intended to perform lip tracking or speech reading. The images were taken in a controlled environment and the subjects were allowed to wear makeup to increase the lip contrast. Additionally, an efficient color segmentation technique was developed to extract lip contour points from color images without manually enhancing the color contrast.

This paper is organized to include the color segmentation technique in Section 2 and lip shape analysis in Section 3. The tongue contact pattern tracking device (the palatometer) and its output are included in Section 4. Section 5 shows both internal and external information and the correlation between them. Conclusions and future work are discussed in Section 6.

2. Shape contour extraction

Images of seven sounds and one resting position reference lip shape were taken for this research. . Active shape model has been used to extract the lip model [6-7]. But, shadows from the lips on the speaker's chin would seriously distort the desired output. In order to accurately extract the lip shapes, color images were taken of the speaker. The lip shapes were extracted using a newly developed lip color segmentation algorithm. We applied a linear color space conversion technique which makes the lip color high value and the skin color low value to perform lip shape segmentation.

The linear color space conversion technique was developed to convert the Red, Green, and Blue (RGB) color space into a one-dimensional (1-D) linear color space. In 3-D RGB color space, the three channels: R, G and B are equally significant in determining a specific color. By converting the 3-D color space into 1-D linear color space, lip shapes can be extracted by selecting a linear threshold to binarize the converted image. We employed the first third-order, which is R×G×B and the full rank of the second-order polynomial to convert the 3-D RGB information into 1-D linear values. The formula is as follows:

$$\text{Linear Color} = C_1 \times R \times G \times B + \sum_{i=2}^{6} C_i \times R^{m_1} \times G^{n_1} \times B^{j_1} + \sum_{i=7}^{10} C_i \times R^{m_2} \times G^{n_2} \times B^{j_2} + C_{11}$$

, where $m_1 + n_1 + j_1 = 2$, $m_2 + n_2 + j_2 = 1$ are all integers. There are 11 coefficients to be determined in this formula.

This algorithm maps every point in an image to a value between 0 and 255 depending on the point's RGB values. In order to calibrate the lip segmentation algorithm, two sets of data were needed. One set contained points that were part of the lips and the other contained points that were not part of the lips. The RGB values of every point in these two data sets were used to calibrate the lip segmentation algorithm. Once the system was calibrated, points that were most similar to the lips mapped to high values while points that were not similar mapped to low values. The output of this system was made into a binary image. Any point greater than the selected threshold turns into a 1 and the rest become zeros. Figure 1 shows the original images of a reference lip (a) and the sound /sh/ (c) and their output binary images from the 1-D color conversion.

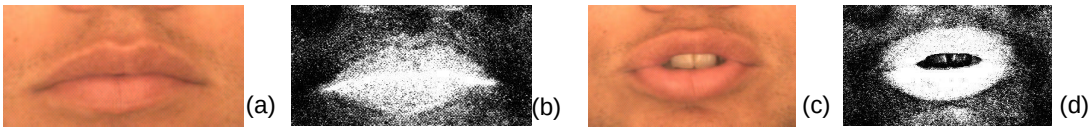

Figure 1. Original images of (a) a reference lip and (c) /sh/ sound and their output of 1-D color space conversion.

Morphological operations were then used to clean up the images. First an opening operation was used which removes from the image all connected components that have fewer than 100 pixels. This operation removed all the small and isolated points that were away from the lips. The remaining feature pixels were then dilated and eroded to remove the rough contour. We then extracted the outlines of the images and smoothed them to remove any sharp edges. The output of the morphological operations and the original images with superimposed lip shapes are shown in Figure 2. Only the outer contour was used for this study. The linear color space conversion provided very reliable color segmentation result for lip contour extraction.

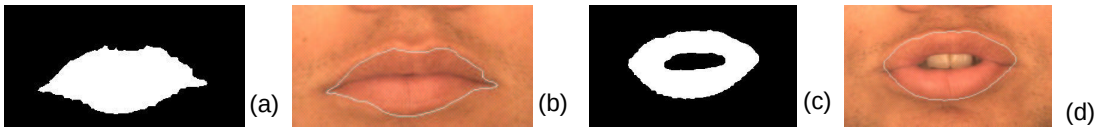

Figure 2. Binary images of (a) a reference lip and (c) the /sh/ sound and their original image with superimposed lip contours

3. Lip shape analysis

A polygon curve representation method was implemented to describe the characteristics of the selected shapes for measuring and quantifying the variations from the reference. This method uses a curve evolution technique to remove small variations and less significant features and then represents the curve in the tangent space, or called turn function [8-10]. This shape representation method works with planar closed curves. While using curve evolution to reduce the number of redundant data points to minimize the computation complexity, this shape representation method preserves some local details so that small variations can be detected. Because images may not always be taken in a fixed position, the shape representations must be invariant to rotation, translation, and scaling transforms so that the small variations between two shapes can be detected.

3.1. Curve evolution

Shape data were extracted from the grayscale image and recorded as a sequence of x and y coordinates as shown in Figure 3. Curve evolution was used to reduce the influence of noise and to simplify the shapes by removing irrelevant and keeping relevant shape features [8-10].

This was achieved by iteratively comparing the relevance measure of all vertices on the polygon. Higher relevance value means that the vertex makes a larger contribution to the shape. The vertex that has the lowest relevance measure was removed and a new segment was established by connecting the two adjacent vertices. The relevance measure was calculated as

$$K(s_1, s_2) = \frac{\beta(s_1, s_2) l(s_1) l(s_2)}{l(s_1) + l(s_2)} \qquad \text{Equation 1}$$

, where β is the turn angle between two adjacent line segments (s_1 and s_2) and $l(s_1)$ and $l(s_2)$ are their normalized lengths. The relevance measure is in direct proportion to the turn angle and the length of the curve segment. A vertex with longer curve segment and/or larger turn angle has a higher relevance measure than the one with shorter length or smaller turn angle. Figure 3 shows the smoothed curves of the target sounds. Each curve has 50 data points preserved for analysis.

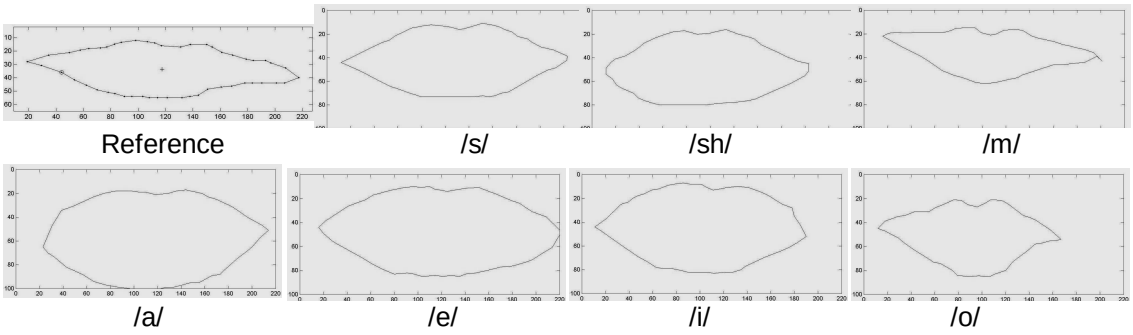

Figure 3. Smoothed lip shapes for study.

3.2. Turn function

The turn function of each shape was represented as a function of normalized length to meet the scaling invariant requirement. Another important requirement for an efficient shape representation method is that the shift on the starting point of the curve should not have any effect on similarity measurement calculations. The turn angle was calculated for each segment by referencing to the horizontal line. The smoothed curve was then represented by the turn function shown in Figure 4. The length was normalized to 1.0 and shown in the x axis. The y axis represents the turn angle. It is translation invariant because the turn angles and length do not contain information about the shape location. It is scaling invariant because it uses normalized length. For rotation and starting point shift, the turn function remains the same shape except shifting vertically when there is a rotation and moving horizontally when there is a shift in starting point.

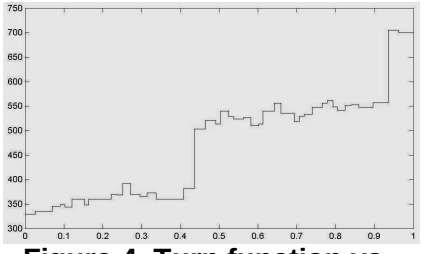

Figure 4. Turn function vs. normalized length.

3.3. Shape variations from reference

Before measuring the distance, the turn functions must be aligned to compensate for the shifts caused by rotation and starting point shift as described in Section 3.2. The distance between two turn functions, Θ_A and Θ_B , can be measured as

$$\delta_2(A, B) = \left\| \Theta_A - \Theta_B \right\|_2 = \sqrt{\left(\int_0^1 \left| \Theta_A - \Theta_B \right|^2 ds \right)} = \sqrt{\min_{\theta \in R, t \in [0,1]} \left(\int_0^1 \left| \Theta_A(s+t) - \Theta_B(s) + \theta \right|^2 ds \right)} \qquad \text{Equation 2}$$

In most cases, the turn functions are not identical because of difference in shape. The alignment can only be achieved through minimizing the distance while shifting one turn function. In other words, the distance between two turn functions is obtained by performing a two-dimensional search to find the minimum distance. Another approach is to reduce the search to one dimension by calculating the best value of the turn angle θ [10]. The best value of θ is a function of length shift t in the x axis to minimize

$$h(t,\theta) = \int_0^1 \left| \Theta_A(s+t) - \Theta_B(s) + \theta \right|^2 ds \qquad \text{Equation 3}$$

when $\theta'(t) = \int_0^1 (\Theta_B(s) - \Theta_A(s+t))\, ds = \alpha - 2\pi t$,where $\alpha = \int_0^1 \Theta_B(s)\, ds - \int_0^1 \Theta_A(s)\, ds$. Equation 4

For each searching step in x (length) direction, the best value of θ was calculated according to Equation 4. The distance δ was calculated and recorded. After shifting the turn function through the searching range, the minimum δ is the distance between the two turn functions. Figure 5 (a) shows the one dimensional searching result for comparing the turn function of sound /a/ to the turn function of reference lip shape. Figure 5 (b) shows the differences between these 2 turn functions for all 50 segments. This variation distribution was divided into 10 regions, 2 for the 2 lip corners and 4 for each of the top and bottom lips for correlation analysis.

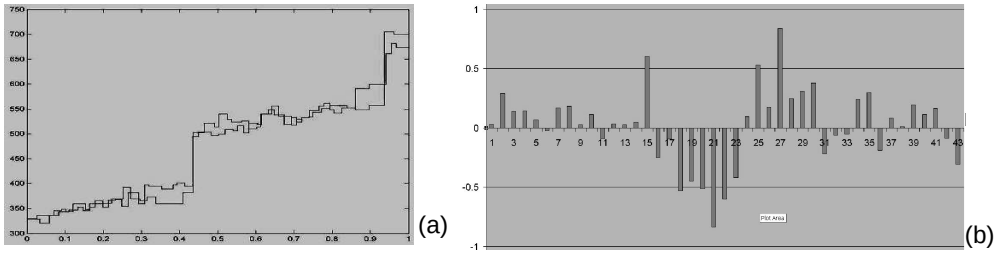

Figure 5. (a) The aligned turn functions and (b) difference in each region

4. Tongue contact patterns

For tongue contact data collection, a palatometer pseudopalate was placed in the subject's mouth. This device consists of a flexible printed circuit (FPC), shown in Figure 6, attached to a custom-made flexible plastic base plate. This device fits snugly against the subject's palate. It detects the contact of the tongue with 118 gold plated electrodes across the surface of the palate to collect information about the position of the tongue during speech. It provides a record of tongue contact patterns 200 times each second, allowing clinically relevant insights into rapid changes in the pattern of the tongue contact during speech.

Figure 6. A complete LogoMetrix's LogoPal.

As shown in Figure 7, sensors are distributed into six regions. When a sensor is contacted during data collection, the dot representing the sensor changes to a filled large circle. It then changes back to a small dot when the contact is released. The essentially instantaneous pattern contrast and the standardized sensor locations allow the recording of changes in articulation place, tongue contact extent, and timing, as well as the subsequent generation of measures that provide quantitative information about articulatory activity.

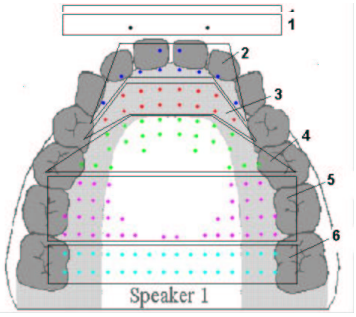

Figure 7. Sensor regions

As shown in Figure 7, there are two sensors in Region 1 for monitoring the lip contact and there are 10, 14, 24, 36, 32 sensors for the remaining regions 2 to 6, respectively. Tongue contact patterns are recorded as number of contacts in each region.

In this research only the static lip shapes and tongue contact patterns were studied. Although the device is capable of recording 200 patterns per second, only one specific pattern that corresponds to one static lip shape was used for analysis. Figure 8 shows the tongue contact patterns of the sounds /s/ and /sh/ and their histogram of each contact region. Each large filled colored circle represents the contact point of the tongue. Small circles represent no contact. Histograms of these two patterns are shown as the percentage of sensors contacted. For example, six out of ten (60%) sensors in region 2 detected tongue contact for the /s/ sound and only 20% for the /sh/ sound.

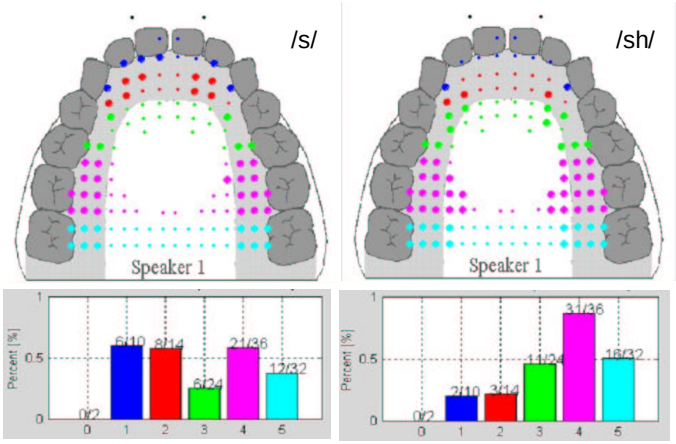

Figure 8. Tongue contact patterns and statistics

5. Correlation

The reference lip shape taken from closed relaxed lips was used as a baseline against which to measure the lip shape variations for each target sound. The variations were measured as distance between two turn functions. The quantified variations, similar to the one shown in Figure 5 (b) for each shape under study, were divided into 10 regions for comparison with the tongue contact data. Figure 9 shows the shape differences in each of the ten lip regions for all three comparisons made. The drop in Region 5 indicates a drop in jaw position when producing the /i/ and /o/ sounds. The high value in Region 6 indicates wider lip shape for producing the /e/ sound. Similarly, the high value in Region 6 indicates wider lip shape for making the /s/ sound.

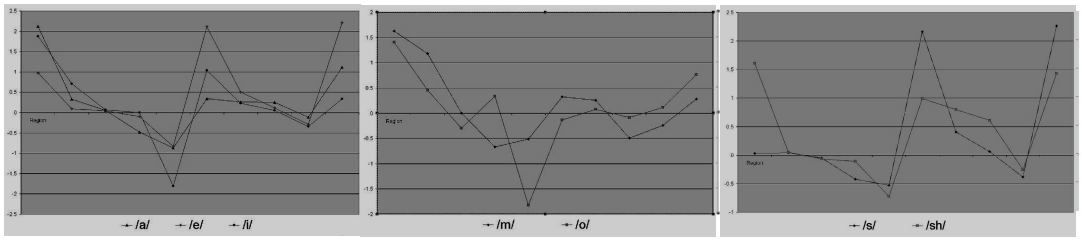

Figure 9. Lip shape comparisons.

The contact pattern statistics data are shown in Table 1. Comparing sounds with similar lip shape, /a/ and /e/, the number of contacts increases in Regions 5 and 6 and also appears in Region 4 as the lip shape widens to change the sound from /a/ to /e/. Contacts in all three regions drop slightly when changing the sound from /e/ to /i/. /o/ and /m/ are two sounds with very different shapes, one is round and the other is elongate. For the /m/ sound, lips are closed and tongue barely makes contact on the edge of Regions 5 and 6, while for /o/, there is no tongue contact at all. The last comparison

Table 1. Contact Pattern Statistics

Region	A	E	I	O	M	S	SH
1	0	0	0	0	100	0	0
2	0	0	0	0	0	60	20
3	0	0	0	0	0	57	21
4	0	29	13	0	0	25	46
5	58	95	89	0	11	58	86
6	38	66	60	0	13	38	50

made in this research was between sounds /s/ and /sh/. For producing the sound /sh/, the tongue must retract slightly to reduce the contact in Regions 1 and 2 and increase the contact in Regions 3, 4, and 5.

6. Conclusions and future work

The palatometer is already being used in clinical settings to help people who are speech impaired. In addition to the valuable data this device can provide, data gathered from visible changes in lip shape may further increase our capacity to help people with speech impairments.

One long-term goal of this line of research, in addition to increasing our understanding of the basic processes of human communication, is to enhance our understanding and treatment of communication disorders. For example, some of the essential characteristics that differentiate individuals who can be well or poorly understood by speech readers have yet to be discovered. Prior studies of speech reading performance have been unable to explain how deaf individuals can glean such a wealth of information from facial cues. The fact that speech reading can be accomplished at all demonstrates that substantial information concerning speech sound formation must be present in visible cues on the face. This research should provide means and methods that may allow discovering specific sources and characteristics of such cues.

Finally, most speech therapy takes pace by having the individual with a speech disorder meet with a speech clinician either individually or as part of a group. Instructions and verbal feedback are provided to guide the client as he or she seeks to imitate the patterns demonstrated by the clinician. Some individuals fail to respond well to this type of imitation-based treatment. In those instances, it would be useful to be able to augment the auditory-visual models with additional instruction based on the type of movement data sought in this research.

7. Acknowledgements

We would like to express our appreciation for the support of LogoMetrix in providing valuable tongue contact pattern information for this research.

8. References

[1] K.K. Neely, "Effect of Visual Factors on the Intelligibility of Speech", I. Acoustical soc. of America, November 1956, vol. 28, no. 6, pp. 1275-1277.

[2] H. McGurk and J. McDonald, Hearing lips and seeing voices, Nature, 265, 1976, pp.746-748

[3] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition", PhD thesis, Univ. of Illinois, Urbana-Champaign, 1984.

[4] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition", Proc. Int'l conf. Acoustics, Speech, and Signal Processing, 1994, pp. 669-672.

[5] T. Chen, R.R. Rao, "Audio-visual integration in multimedia communication," Proceedings of the IEEE *International Conference on Acoustics, Speech, and Signal Processing,1997*, volume 1, pp. 179-182.

[6] K.L. Sum, W.H. Lau, S.H. Leung, A.W.C. Liew, and K.W. Tse, "A new optimization procedure for extracting the point-based lip contour using active shape model", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, May 2001, vol. 3, pp. 1485–1488.

[7] A. Caplier, "Lip detection and tracking", Proceedings of IEEE 11th International Conference on Image Analysis and Processing , Palermo, Italy, September 2001, pp. 8-13.

[8] L.J. Latecki and R. Lakämper, "Application Of Planar Shape Comparison To Object Retrieval In Image Databases". Pattern Recognition, 2002, vol. 35, no, pp. 15-29.

[9] L.J. Latecki and R. Lakämper, "Shape Description and Search for Similar Objects in Image Databases", State-of-the-Art in Content-Based Image and Video Retrieval, Kluwer Academic Publishers, 2001.

[10] E. M. Arkin, L. Paul Chew, D. P. Huttenlocher, K. Kedem, and J. S. B. Mitchell, "An Efficient Computable Metric for Comparing Polygon Shapes", IEEE Transactions on Pattern Analysis and Machine Intelligence, March 1991, vol. 13, no. 3, pp. 209-216.